White Paper

Report ID: 115807

Application Number: HG-229309-15

Project Director: Brian James MacWhinney

Institution: Carnegie Mellon University

Reporting Period: 4/1/2015-3/31/2017

Report Due: 6/30/2017

Date Submitted: 6/7/2017

Grant Number HG-229309-15
LangBank: Digital Infrastructure to Support the Study of Latin and Historical German
Project Director: Brian MacWhinney
Grantee Institution: Carnegie Mellon University
Date: June 7, 2017

This project is designed to create web-accessible materials for the study of Classical Latin. These materials encode each of the 750 texts in the Perseus database in the JSON computer format to allow for web-based display of word meanings, morphosyntactic information, literal English translation, and alignment with free translations. Our audience for this work includes scholars and university learners of Latin internationally. We have also begun outreach to High School students learning Latin in the United States.

Using materials from several treebanks, we have trained and tagged the text of Caesar's *Gallic Wars* and used this training set to tag new texts. To accomplish this, it was necessary to harmonize part of speech and grammatical dependency tags between existing datasets to bring them into accord with each other and the requirements of the TalkBank software. The resultant tagged texts are now compatible with the CLAN software that is used for other TalkBank projects (http://talkbank.org) and can also be displayed over the web using TalkBank software. The tagging software we are using is the MALT dependency parser, which has been used successfully in a variety of current NLP (natural language processing) applications.

We have also developed a method for learning Latin grammatical structures through filling in of cloze items in Wikipedia pages. We had first developed this method for learning how to select the correct form of the definite article in German. For Latin, this application involves selection of the correct case form of the noun.

Our currently available materials are online at http://sla.talkbank.org/latin/ These include:

- JSON Viewer Demo
- Multiple Document Demo
- Perseus ALigment Demo
- Database structure documentation: example JSON document, TEI tags in Perseus
- Latin Wikipedia cloze exercise

We also developed a method for importing CHAT data into the ANNIS corpus tool system developed by our colleagues in Berlin. This will allow for ongoing integration of the Latin and German materials.

This work confronted three problems. The first is that, after subtraction of indirect costs by the University, the level of funding was not enough to support a full-time programmer. The second was that the Perseus database upon which we needed to rely had serious gaps in its structure, forcing us to essentially redo years of work done by the Perseids and Alpheios projects. The third problem was that there was a lag in funding startup between the NEH and DFG projects because of restrictions in DFG funding documentation.

Of these three problems, the second was the most serious. We have managed to overcome much of this problem through formulation of a new JSON structure for Latin texts. However that work is not yet complete. We plan to complete the JSON restructuring of the Perseus database after the end of NEH funding by relying on funding from other sources. Once the work on completion of the JSON restructuring of Perseus is completed, we believe that the project will have a major impact on work in digital studies of Latin.